Xeon+FPGA Platform for the Data Center

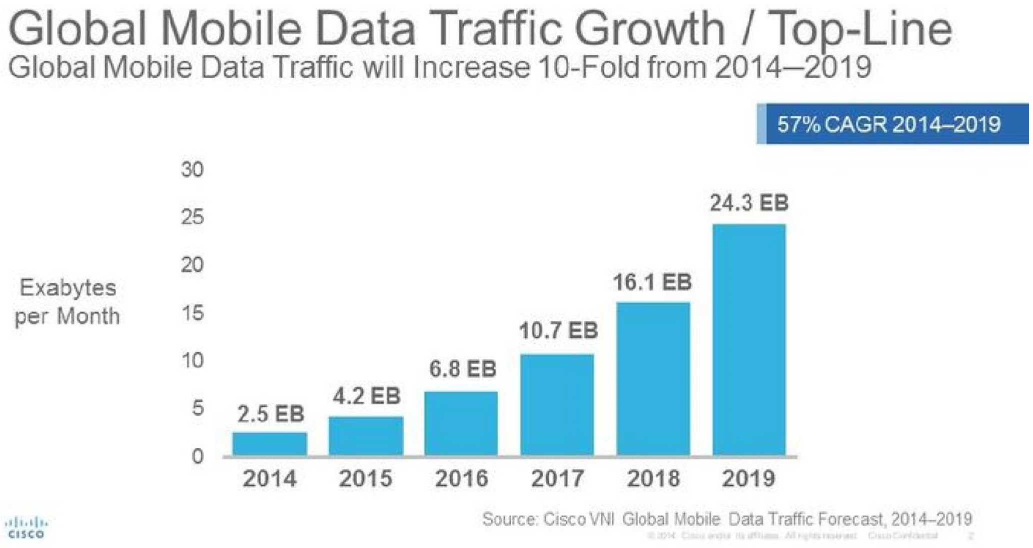
ISCA/CARL 2015

PK Gupta, Director of Cloud Platform Technology, DCG/CPG

Overview

- Data Center and Workloads
- Xeon+FPGA Accelerator Platform
- Applications and Eco-system

Exponential growth in mobile....



Global Mobile Data Traffic Growth / Top-Line

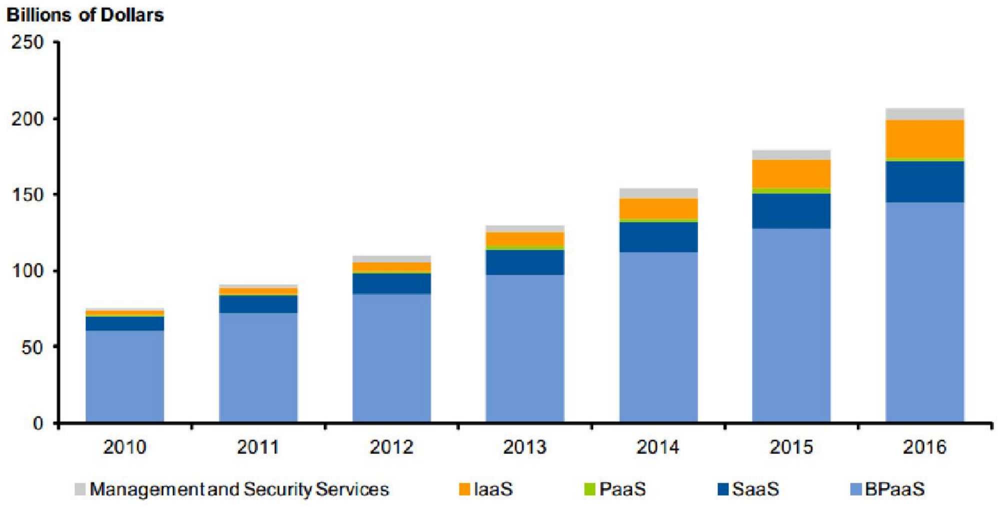
Global Mobile Data Traffic will Increase 10-Fold from 2014–2019

57% CAGR 2014–2019

Source: Cisco VNI Global Mobile Data Traffic Forecast, 2014–2019

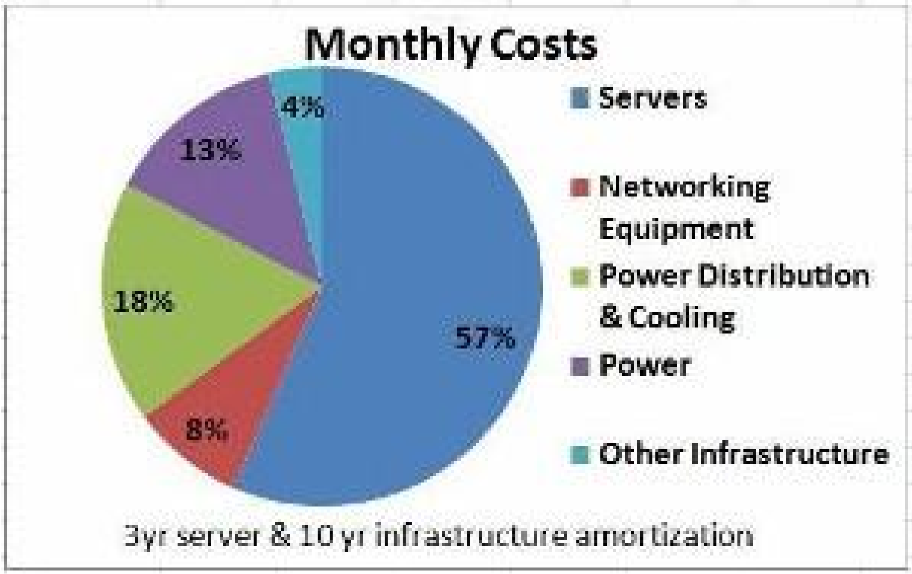
…is driving Data Center growth

Figure 3. Public Cloud Services Market Size by Segment, 2010-2016



Source: Gartner (August 2012)

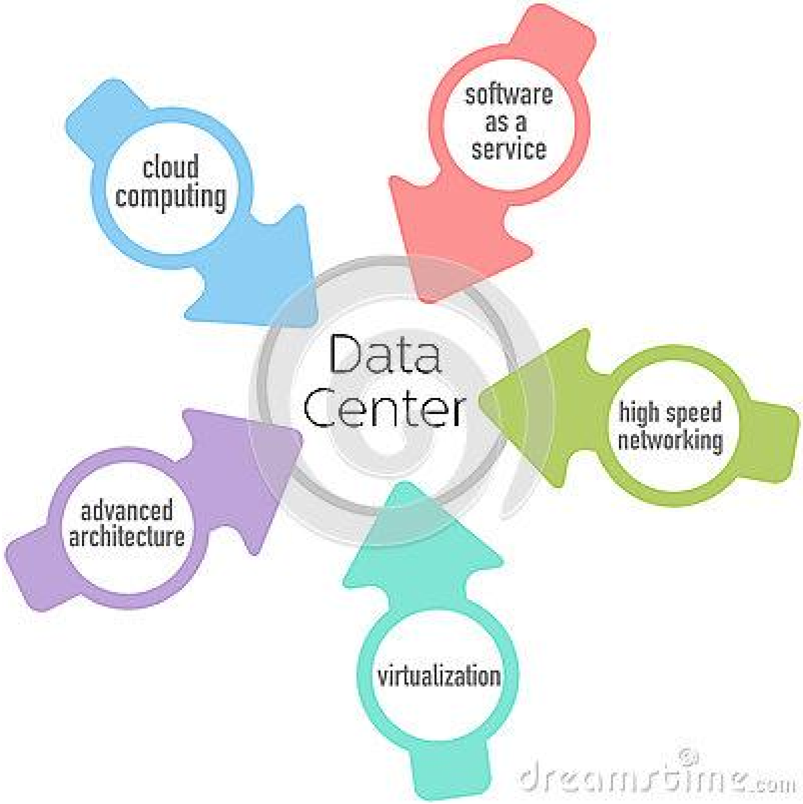
Leading to search for greater performance efficiencies…



…Across Data Center Workloads

- Diverse workloads:
  - Cloud Services: Search, Web Servers, ..
  - Analytics: Big Data, Machine Learning, …
  - Scientific: Genomics, Security, …
  - Communication: Packet Processing, Virtual Switching, …
  - Storage: Compression, Deduplication, …
- Changing dynamics:
  - No single killer app
  - Emerging new apps drive changes in workloads

A homogenous compute platform for the Data Center?



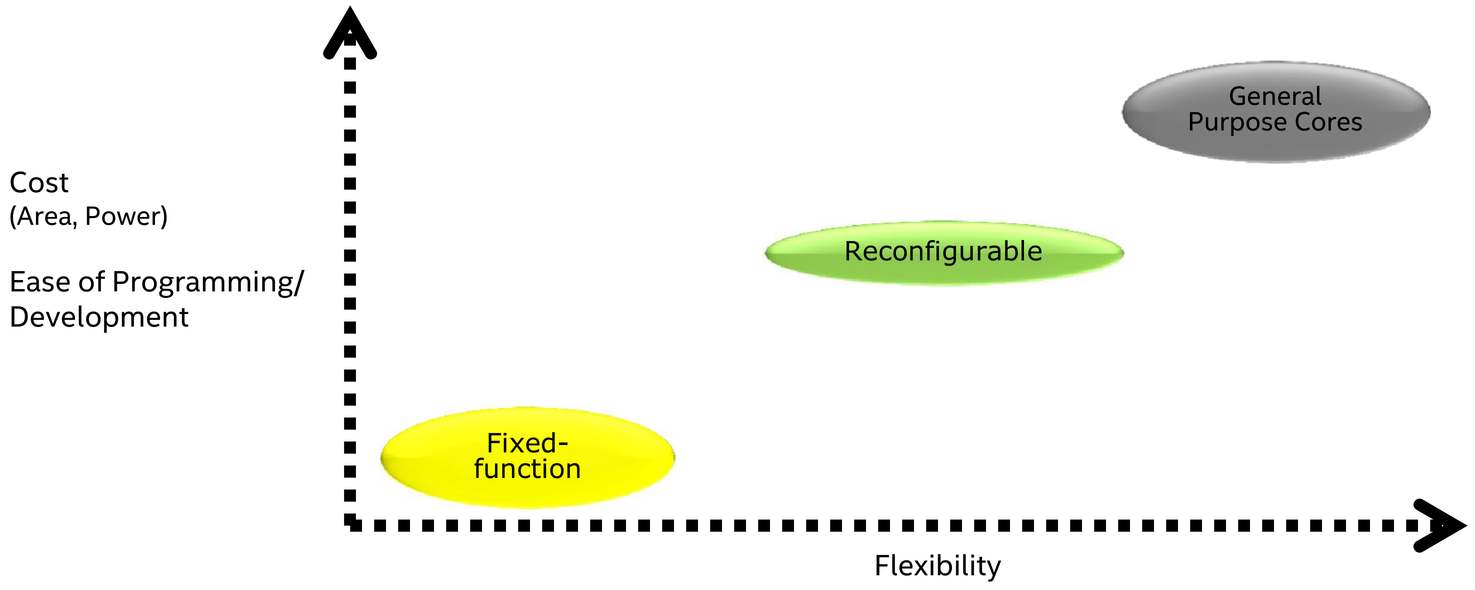
Overview

- Data Center and Workloads
- Xeon+FPGA Accelerator Platform
- Applications and Eco-system

Motivation for Accelerators

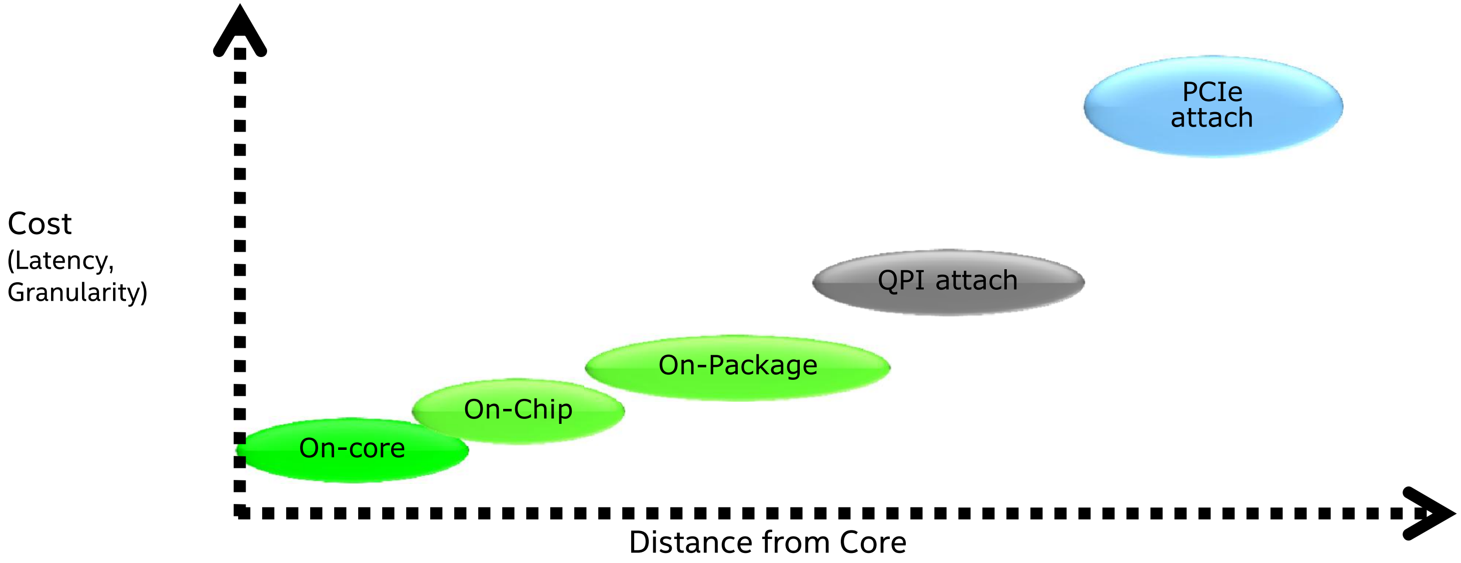
- **Enhanced Performance**: Accelerators compliment CPU cores to meet market needs for performance of diverse workloads in the Data Center:
  - Enhance single thread performance with tightly coupled accelerators or compliment multi-core performance with loosely coupled accelerators via PCIe or QPI attach
- **Move to Heterogeneous Computing**: Moore's Law continues but demands radical changes in architecture and software.
  - Architectures will go beyond homogeneous parallelism, embrace heterogeneity, and exploit the bounty of transistors to incorporate application-customized hardware.

Accelerator Architecture



Performance Efficiency: Performance/Watt, Performance/$
Programming Complexity : Effort, Cost

Accelerator Attach



Best attach technology might be application or even algorithm dependent

Coherency and Programming Model

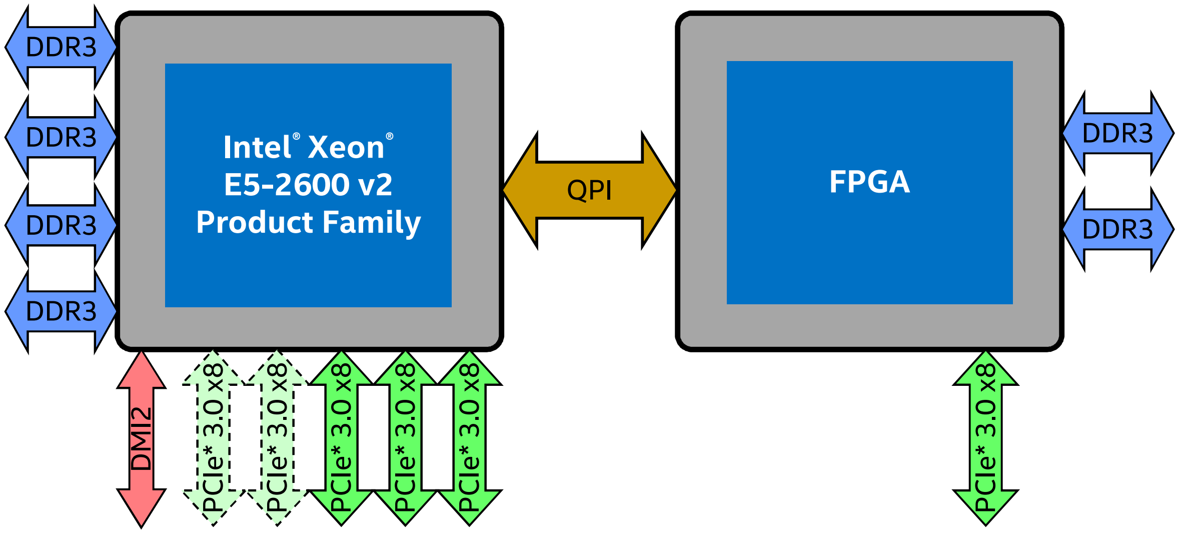
- Data Movement
  - In-line
    - Accelerator processes data fully or partially from direct I/O
  - Shared Virtual Memory :
    - Virtual addressing eliminates need for pinning memory buffers
    - Zero-copy data buffers
- Interaction between Core and Accelerator
  - Off-load
  - Hybrid : algorithm implemented on host and accelerator

Proposed Platform for the Data Center

- FPGA with coherent low-latency interconnect:
  - Simplified programming model
    - Support for virtual addressing
    - Data Caching
  - Enables new classes of algorithms for acceleration with:
    - Full access to system memory
    - Support for efficient irregular data pattern access
  - Remapping of algorithms from off-load model to hybrid processing model
    - Fine grained interactions

IVB+FPGA Software Development Platform

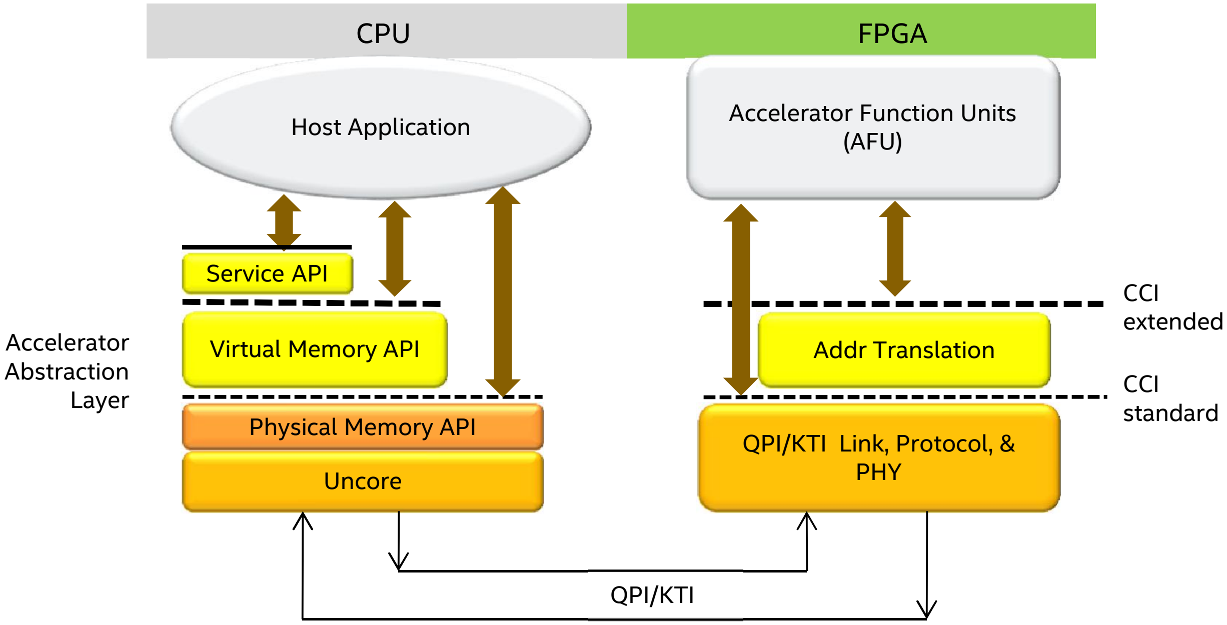
Software Development for Accelerating Workloads using Xeon and coherently attached FPGA in-socket



Processor	Intel® Xeon® E5-26xx v2 Processor
FPGA Module	Altera Stratix V
QPI Speed	6.4 GT/s full width (target 8.0 GT/s at full width)
Memory to FPGA Module	2 channels of DDR3 (up to 64 GB)
Expansion connector to FPGA Module	PCIe 3.0 x8 lanes - maybe used for direct I/O e.g. Ethernet
Features	Configuration Agent, Caching Agent,, (optional) Memory Controller
Software	Accelerator Abstraction Layer (AAL) runtime, drivers, sample applications

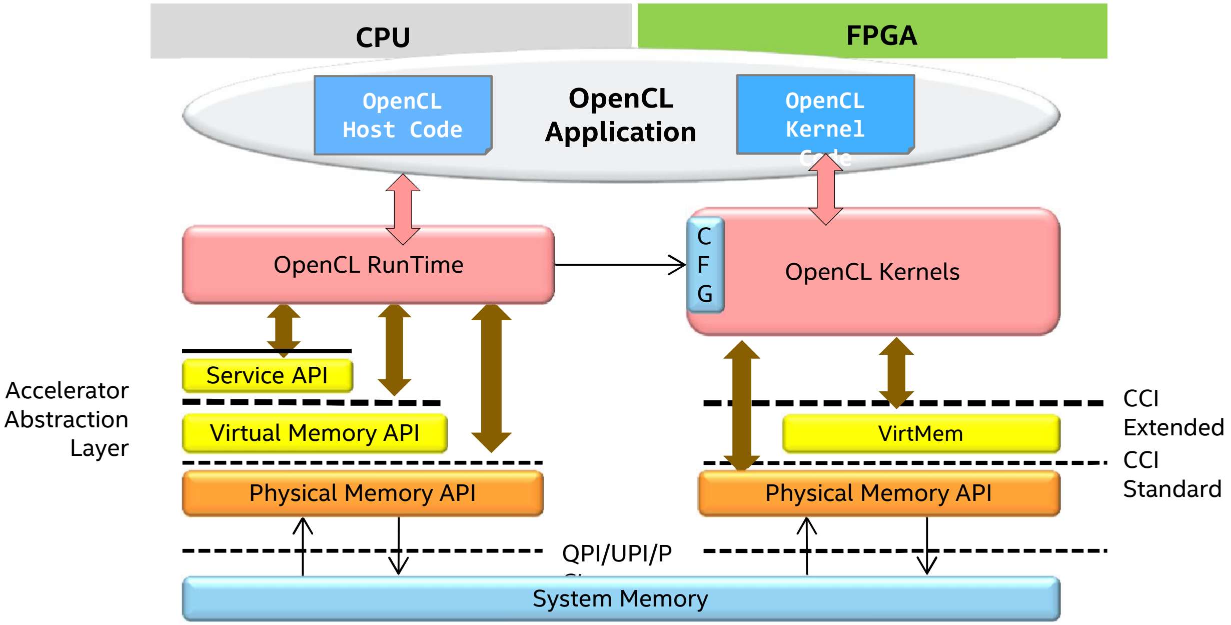
Heterogeneous architecture with homogenous platform support

Programming Interfaces



Programming interfaces will be forward compatible from SDP to future MCP solutions
Simulation Environment available for development of SW and RTL

Programming Interfaces : OpenCL

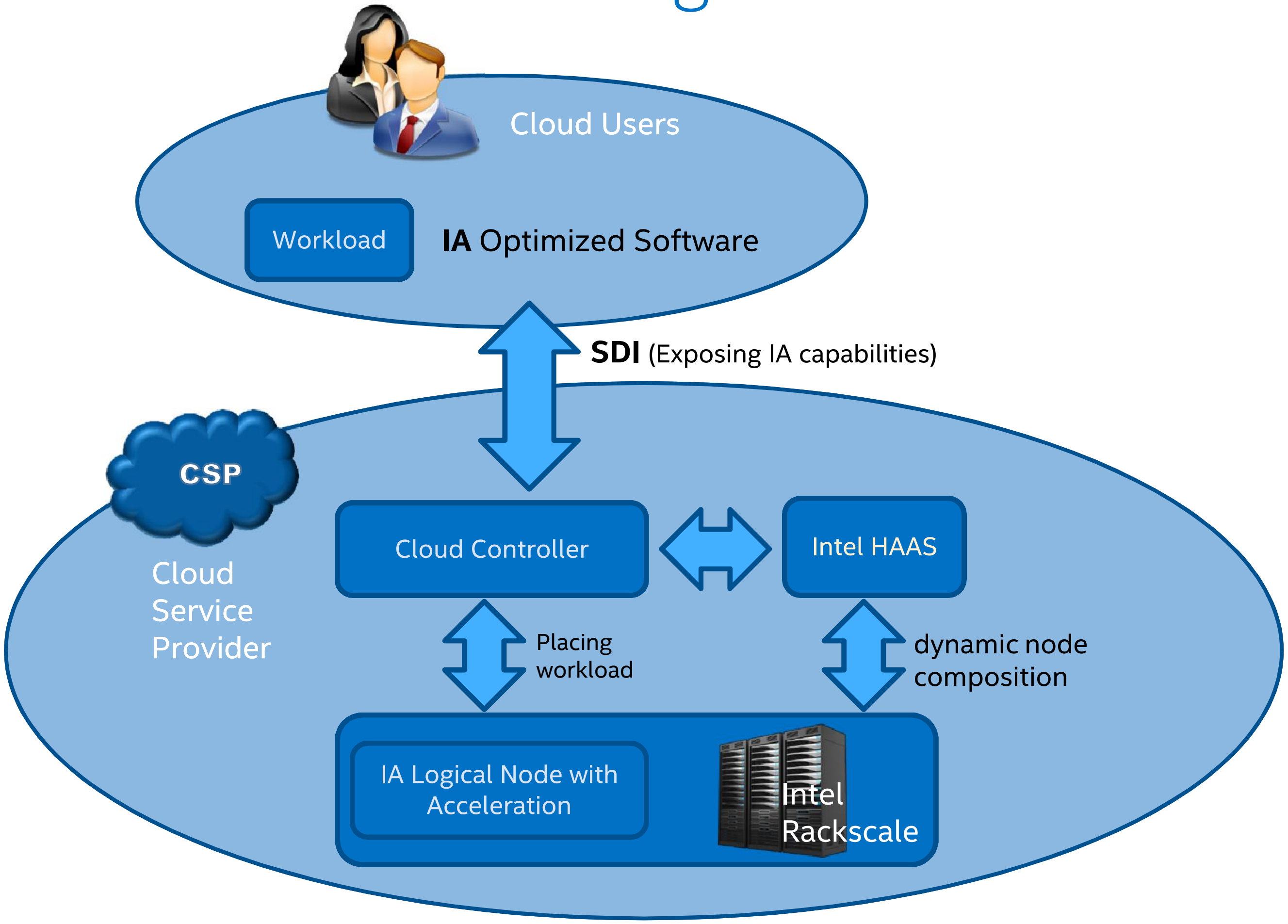


Unified application code abstracted from the hardware environment
Portable across generations and families of CPUs and FPGAs

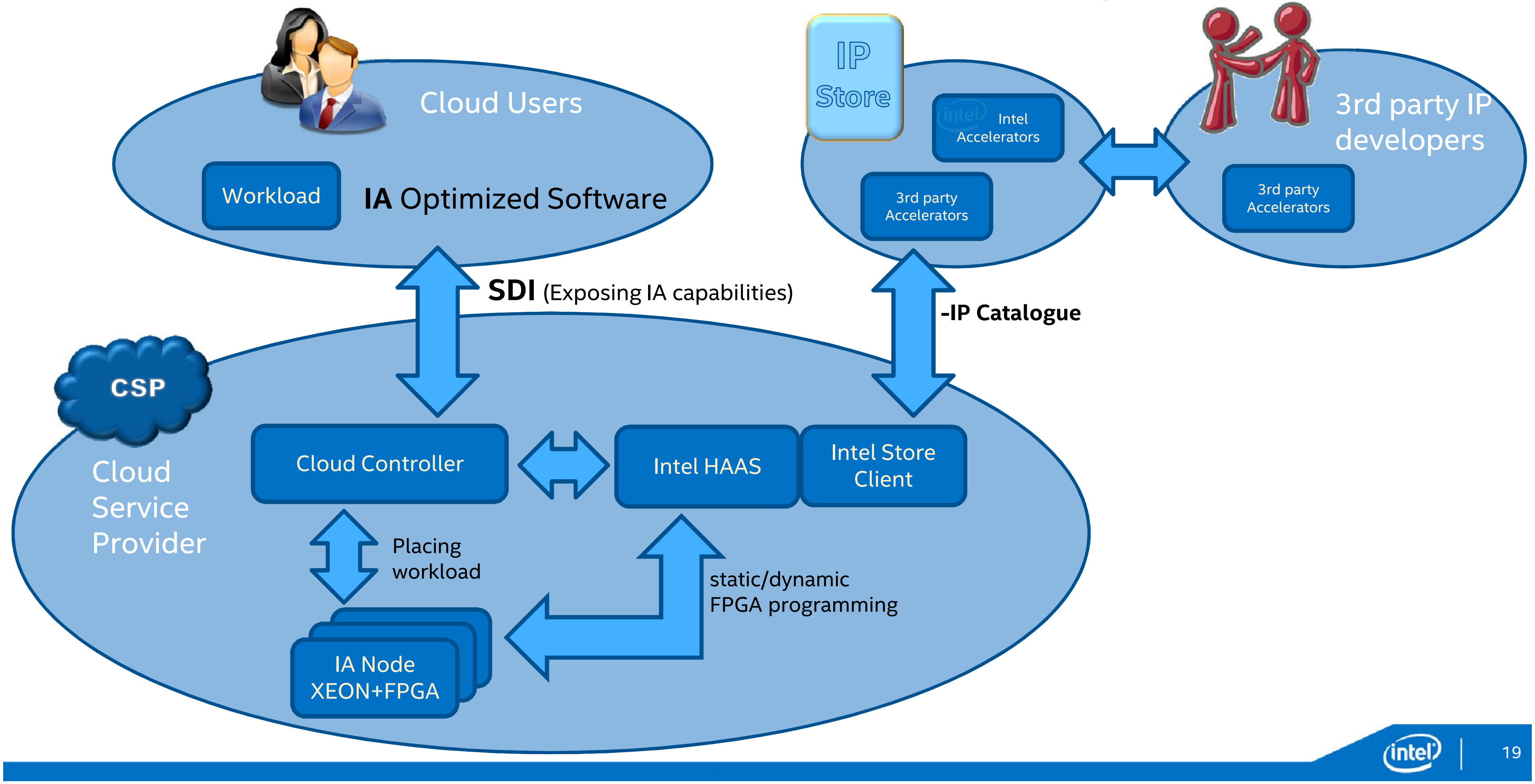
Overview

- Data Center and Workloads
- Xeon+FPGA Accelerator Platform
- Applications and Eco-system

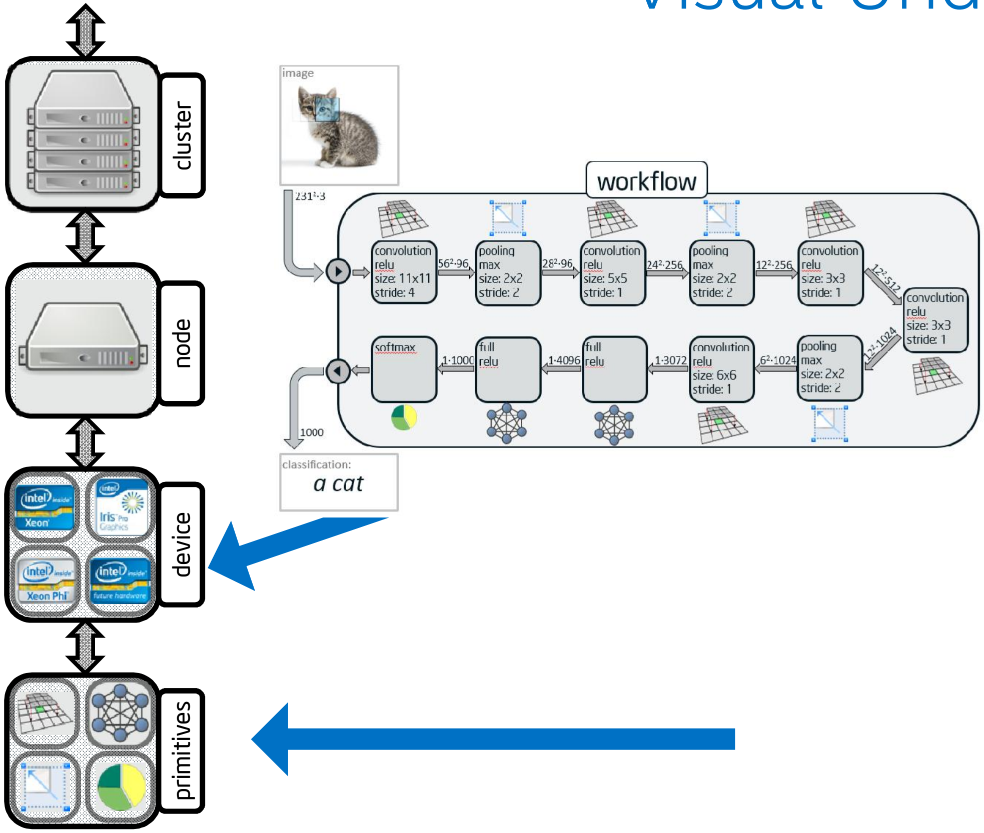
XEON+FPGA in the Cloud : integration with SDI and RSA

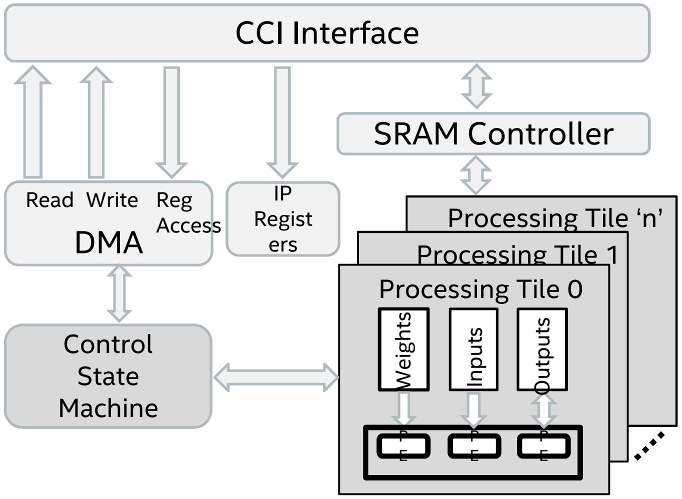


XEON+FPGA in the Cloud: IP Store Concept

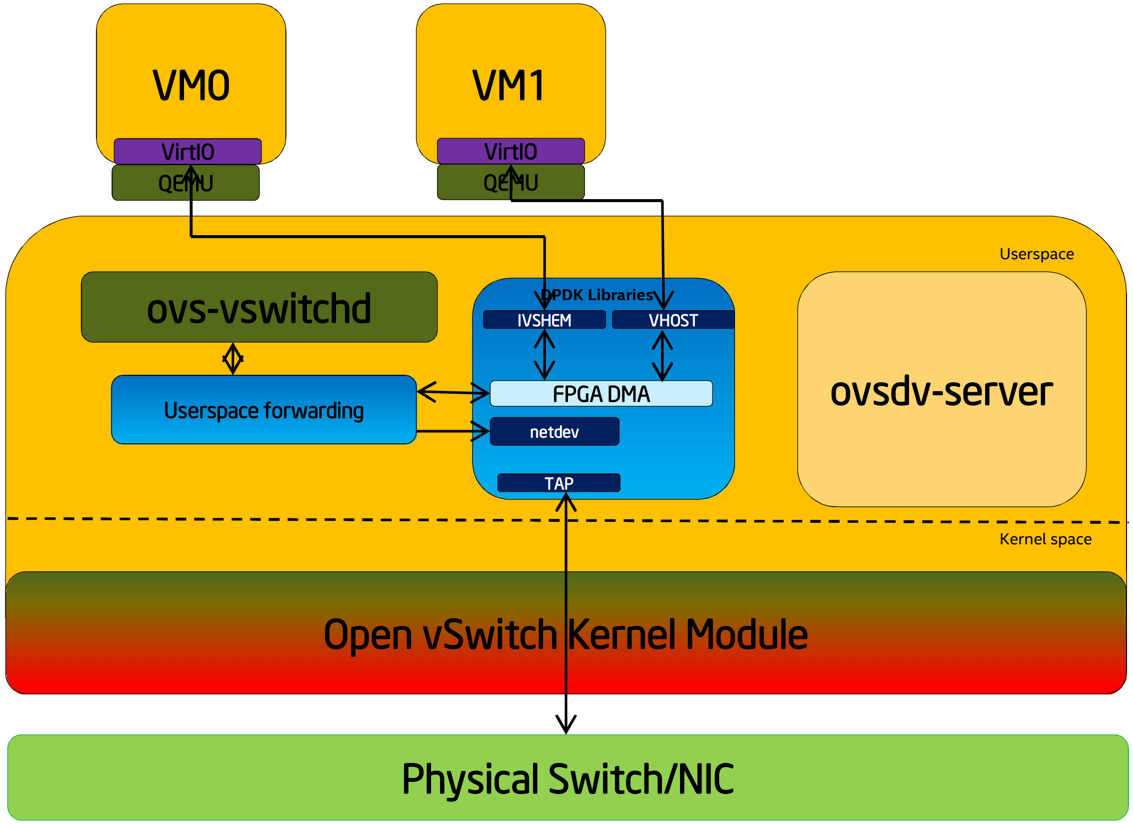


Example Usage : Deep Learning Framework for Visual Understanding



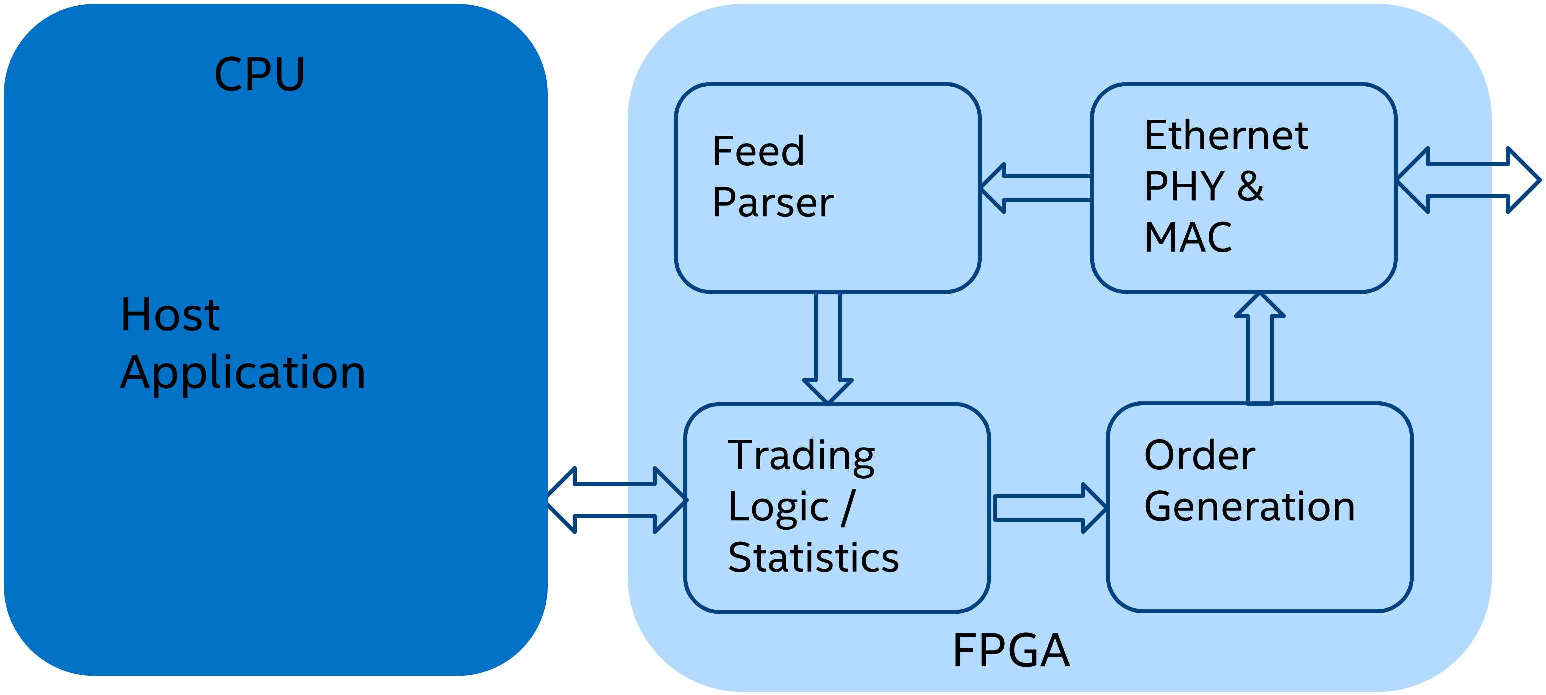


Example Usage: Accelerating Open VSwitch w/DPDK

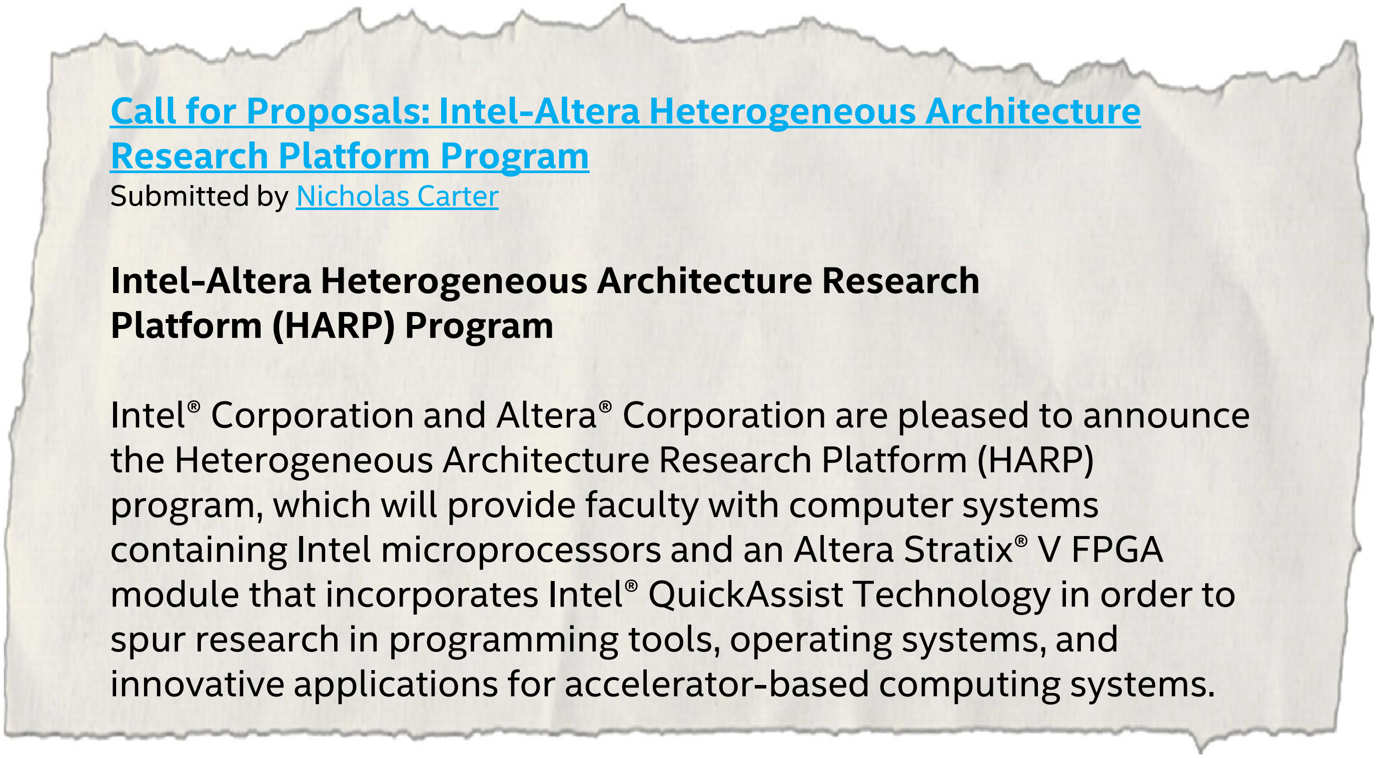


- Offload DMA Engine to FPGA :
  - Frees up CPU cycles to perform more useful work
  - Reduce cache pollution.
- Add support for Packet Classification, ACL, and other functions including Direct I/O in FPGA

Example Usage: High Frequency Trading Accelerator



Academic Research



Call for Proposals: Intel-Altera Heterogeneous Architecture Research Platform Program

Submitted by Nicholas Carter

Intel-Altera Heterogeneous Architecture Research Platform (HARP) Program

Intel® Corporation and Altera® Corporation are pleased to announce the Heterogeneous Architecture Research Platform (HARP) program, which will provide faculty with computer systems containing Intel microprocessors and an Altera Stratix® V FPGA module that incorporates Intel® QuickAssist Technology in order to spur research in programming tools, operating systems, and innovative applications for accelerator-based computing systems.

Q & A